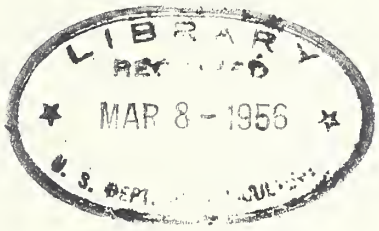
UNITED STATES DEPARTMENT OF AGRICULTURE
AGRICULTURAL RESEARCH SERVICE
Washington, D.C

The Problem of Unequal Variances

by

Glenn L. Burrows
Agricultural Marketing Service



Sponsored by
Departmental Committee on Experimental Design
1954

Copies of this paper can be obtained upon request from:

Biometrical Services
Office of Administrator
Agricultural Research Service
Washington 25, D.C

The following lectures have been sponsored by the Departmental Committee on Experimental Design and copies may be obtained from the Biometrical Services, Office of the Administrator, Agricultural Research Service, Washington 25, D. C.

May 16, 1951 — The Application of Punch Card Equipment to Statistical Processing.
Lawrence Armstrong, Assistant Chief
Machine Tabulation Division, Bureau of the Census
United States Department of Commerce

Feb. 6, 1952 — Determination of the Size of Experiments and Samples.
William G. Cochran, The Johns Hopkins University

Jan.11, 1951 — The Value and Usefulness of Statistics in Research.
Prof. Gertrude M. Cox, Director, Institute of Statistics
University of North Carolina, Raleigh, North Carolina

Mar.30, 1953 — The Value of Summarizing Results From Individual Experiments.
Walter T. Federer, Professor of Biological Statistics
Department of Plant Breeding, Cornell University
Ithaca, New York

Feb.26, 1953 — Some Lattice Designs.
Boyd Harshbarger, Virginia Agricultural Experiment
Station of the Virginia Polytechnic Institute

Feb.15, 1950 — Sampling Methods in Marketing Research.
Earl E. Houseman, Statistical Consultant
Bureau of Agricultural Economics

Dec.12, 1951 — A Review of the Principles of Experimental Design and Some Applications.
Professor Oscar Kempthorne, Professor of Statistics
Statistical Laboratory, Iowa State College, Ames, Iowa

Sept.6, 1950 — Statistics and Research on Pasture and Grazing.
Dr. Henry L. Lucas, Animal Science Statistician
Department of Experimental Statistics
North Carolina State College, Raleigh, North Carolina

Dec.14, 1949 — How Statistics Improves Physical, Chemical, and Engineering Measurements.
Dr. William J. Youden, Assistant Chief
Statistical Engineering Section, National Bureau of Standards

29 Sept 54

The Problem of Unequal Variances

The title, "The Problem of Unequal Variances," is too broad; far too little is known about many facets of the problem to attempt to discuss it generally. Rather we will consider here only solutions to certain aspects of the problem that have been published either quite recently or in such form as to gain popular acceptance slowly. It is part of our responsibility to see that these improved techniques become known.

There isn't one of us here today who wasn't warned in his first course in statistics against the so-called L.S.D. (or "least significant difference") test for the separation of observed sample means, and for most of us that first course was quite some years ago. It was over 15 years ago that Fisher suggested his now crude technique for overcoming some of the bias in that test. However, at least since Tukey's 1949 Biometrics paper followed by the work of Duncan, Sheffe, Bechhofer, and others, the beast should long since have succumbed. Nevertheless, many of us see almost daily evidence of the continued use of the "old favorite" L.S.D.

Probably the most frequently made decision in experimental science is whether one mean is greater than one or more other means; that is, whether on the basis of sample evidence certain population means may be supposed with any reasonable assurance to differ. To aid in making such a decision, there are several procedures available more or less sensitive to the quantity of other available and pertinent information such for example as the knowledge of the magnitude of the population variances or of the equality of the population variances or such as a set of sample estimates of the population variances possibly or only feel for the range of possible values. There is no one, I suppose, who, when confronted with sample estimates of population means, does not temper his judgment as to the comparative magnitudes of population means by some outside estimate of precision of these means. It may be by intuition, by his faith in the experimentor, by his own past experience or by some evidence contained in the samples themselves. But the precision of the sample means and therefor of the differences among them is questioned. For example, we are told that samples from two lots of wool yielded average fiber diameters of 21.40 microns and 31.80 microns respectively. Is it reasonable to suppose that the mean fiber diameters of the two lots differ? We are further told that experienced wool graders selected the samples; is the question any easier to answer? We add to our information the knowledge that the same experienced graders on the basis of these samples, graded the two lots respectively 64s stable and 48/46s staple. At this point, any experienced grader would no doubt conclude that the lot mean fiber diameters differed, and he would probably arrive at this conclusion not so much on the basis of the measurements as upon his knowledge, from experience, of the ability of graders to detect fiber diameter differences despite the usual variability among samples and among fibers in samples. Possibly, even the knowledge that both lots were graded staple length assists him in appraising the precision with which the graders are able to separate lots of different mean fiber diameters on the basis of samples from those lots. It must be remembered that a fiber, however small, is after all not uniformly cylindrical throughout, and, just as for their human counterparts, "diameter" may be considerably easier to agree upon for the svelte type than for the buxom

model. At least an experienced grader is probably well aware of the relationship exhibited in Figure 1, although he might not recognize the language by which we describe it. However, we do not have the benefit of his experience, and even the additional information contained in Figure 1 is still insufficient to answer our question. Not only that, the knowledge from Figure 1 strikes terror to our hearts, but only because we're statisticians. We all knew, of course, as soon as we had the standard deviations, or estimates thereof, that the only missing links were the sample sizes. But now, even the knowledge of the sample sizes won't close the gap because the usual tests are valid only if we can assume equal variances.....and this not only on the null but also on the non null hypothesis. Incidentally, I mention this because Figure 1 was plotted wholly from staple length measurements; for other fiber lengths the standard deviations may be quite different. Thus even on the null hypothesis of equal mean fiber diameters there is no assurance of equal standard deviations. The samples consisted of 600 and 1000 fibers respectively; so the standard errors of the means are respectively: $\frac{4.63}{\sqrt{600}} = .189$ and $\frac{7.25}{\sqrt{1000}} = .229$. And so my example explodes! No one could possibly doubt the significance of the difference of over 10 microns between sample means with standard errors of such magnitude; neither should anyone take so many observations to establish the significance of such a difference. The truth of the matter is the experiment from which these data came was designed "(a) to determine if the size of coring tube used in drawing cores influenced fineness results, (b) to develop a reliable and economical plan for sub sampling scoured core residues, and (c) to determine the adequacy of cores drawn for clean yield for estimating the fineness and variability of a lot of graded grease wool."1/ A careful reading of the manuscript cited suggests that: We may interpret (a) to mean: Is there a significant difference between the mean fiber diameters for the composite 3/8" and the composite 1-1/4" coring methods lot by lot? (See Table 1), (b) involves answering the following question: Is there a significant difference between plans of sub sampling I and II? It also involves answering the question: Do the mean fiber diameters for the "Individual" and "Composite" methods of core sampling differ significantly? We may interpret (c) to mean: Do any of the mean fiber diameters for the various core sampling procedures differ significantly from the mean for the same lot as determined from the card sliver? Supplementally, but not alternatively, the authors may have been interested in whether corresponding differences among standard deviations and among coefficients of variation could be shown to be significant; possibly they may even have been interested in the significance of the differences from lot to lot among coefficients of variation for the same coring method. So the example does, after all, exhibit the two usual aspects of the problem of unequal variances, namely; are the variances homogeneous and if so, how do we test the differences between means? Thus, for the 48/46s, it is desired among other things to know whether the individual 3/8" core sampling method yielded a mean fiber diameter significantly different from that of the card sliver sample. In light of our condemnation of the L.S.D. test and certainly from the standpoint of gaining the greatest sensitivity, we should view our objectives collectively. Having already disposed of the question of equality

1/ Core-sampling Grease Wool for Fineness and Variability, D. D. Johnston, W. J. Manning, H. D. Ray, W. A. Mueller, and E.M. Pohle, USDA. In press.

of means and variances between different lots in the negative, suppose we concentrate solely upon the 48/46s and ask first whether the sampling method standard deviations differ, not just for the individual 3/8" and card sliver samples, but collectively.

For this purpose we use the familiar Bartlett test[2/]:

$$M = n \ln \left(\frac{\Sigma n_i s_i^2}{n} \right) - \Sigma n_i \ln s_i^2$$

where ln indicates logarithm to the base e and

$$n = \Sigma n_i$$

In our example $n_i \doteq 1000$ all i; we will assume $n_i = 1000$ for all i.

$$\text{then } M = 1000 \left\{ 6 \ln \frac{\Sigma s_i^2}{6} - \Sigma \ln s_i^2 \right\}$$

$$= 1000 \left\{ 6 \ln \frac{346.5}{6} - 2 \Sigma \ln s_i \right\} = 1000 \left\{ 6 \ln \frac{346.5}{6} - 2(12.16160) \right\}$$

$$= 1000 \left\{ 24.33654 - 24.32320 \right\}$$

$$= 13.34$$

$$M' = \frac{M}{1 + \frac{\Sigma \frac{1}{n_i} - \frac{1}{n}}{3(k-1)}}, \text{ where k is the number of variances tested.}$$

In our example: $M' = \frac{13.34}{1.0004} = 13.33$

Comparing this with the tabular probabilities for χ^2 with k-1 = 5 d.f. we find that the probability is only about .02 of exceeding such a value by chance if the variances were in fact equal. At least one must reject the hypothesis of equal standard deviations at the 5 percent level of significance. How many of you would have predicted this having merely looked at the data in table 1?

There was, of course, on the collective null hypothesis of equal means and variances still another estimate of the standard deviation, namely, that

---

2/ For a wholly readable discussion of Bartlett's test see Rao, Advanced Statistical Methods in Biometric Research pp. 226-230.

given by the sum of squares of the means about their own mean. Thus

$$s_m^2 = \left\{ \Sigma m_i^2 \quad \frac{(\Sigma m_i)^2}{k} \right\} /k-1$$

$$= .05018,$$

or on a per fiber basis $s^2 = 50.18$ and $s = 7.08$.

It is now clear that had we known nothing about equality of means or variances and wished to test the collective null hypothesis we might have done so exactly as above except for the inclusion of one more estimate of variance. The new M is easily computed

$$M = 6005 \ln \left\{ \frac{346,743.8}{6005} \right\} - 24342.77$$

$$= 24356.28 - 24342.77 = 13.51$$

and $$M' = \frac{13.51}{1.0114} = 13.36$$

As was to be expected, M has changed little because, even though the new s is less than all the other 6, it carries only 5 d.f. compared with 1000 for each of the other 6. The new M exceeds the 5 percent tabular χ^2 at 6 d.f.; so we conclude at this level of significance that the collective null hypothesis is false. That is, either the means are unequal, the standard deviations (variances) are unequal or both. This test did not tell us why the collective hypothesis is false, only that it is false. It is easy to see in this instance that the significance of M is due to lack of homogeneity of variances rather than to inequality of means; the estimate of variance derived from the among-methods mean square was less than all other within methods estimates. It is also easy to see, however, that the above test would be relatively insensitive to differences between means because of the few degrees of freedom for the mean square among method means.

Under the collective null hypothesis, but considering non-null hypotheses that permit only unequal means, not unequal variances, we would have used the familiar analysis of variance F test as our criterion. We have shown by our first application of Bartlett's test that the assumptions required for the analysis of variance test are probably invalid. We could, of course, use the analysis of variance test and place our trust in those few eminent statisticians who have dared assert that such a test is probably not too badly biased. But, frankly, I fear their audacity and their eminence would be our only refuge, for, the more I probe, the less factually corroborative evidence I find to support their opinion in general. G. E. P. Box, in the June 1954

Annals of Mathematical Statistics summarizes some of the recent constructive contributions as follows:

"The problem of the effect of unequal group variances was considered in the case of the t test by Welch (5). He obtained approximate probabilities from which it appeared that the effect was small when the groups were of equal size, but larger when they were different in size. Later some exact probabilities for this case were found by Hsu (9) and another investigation by a different approximate method was made by Grunow (10). Both of these investigations confirmed Welch's results. Quensel (11) considered the one-way analysis of variance classification more generally and obtained an approximate expression for the variance of the F criterion when the group variances differed. He concluded that the test would not be greatly affected if the group sizes were equal.

David and Johnson (12), (13), (14) have discussed the general problem of the power function of analysis of variance criteria when the observations are distributed independently but do not necessarily follow the normal distribution or have constant variance. As a special case, they consider the one-way classification in which the observations are normally distributed but the variances differ from group to group. Their method is different from that given here and is an approximate one. At the time of writing, they have published few numerical results and these (14) are confined to the case in which the sizes of the groups are all equal. Confirming the results of Quensel, only slight changes in probability, from those expected if the assumptions were true, have been found."

On the basis of these works and his own research, he concludes as follows, concerning the effect of group-to-group inequality of variance in a one way classification:

> "It appears that if groups are equal, moderate inequality of variance does not seriously affect the test. However, with unequal groups, much larger discrepancies appear."

To be sure in our immediate example, all means are based upon equal numbers of fiber measurements, but this is far from true in the case of the lot of 64s. You will, no doubt, call my attention to the sizable number of degrees of freedom available for estimating the standard deviations. Why not assume that $\Sigma \left(\frac{(m_i - \bar{m})^2}{s_i^2/n_i} \right)$ is distributed as χ^2 with k-1 = 5 d.f.; where, of course, $\bar{m}$ is not $\frac{\Sigma m_i}{n}$ but rather $\frac{1}{n} \Sigma m_i/(s_i^2/n_i)$? And with this question, my example explodes completely!

The full truth of the matter is that I deliberately selected an example wherein I could dodge the issue for the simple reason that this is our usual tactic. But no one here is deluded by such trickery. We can in our real every day life rarely avoid the unequal subclass numbers problem or the

problem of unequal variances either by the usual tricks you find in the text books or by exploding the sample sizes. Until quite recently we have had no satisfactory techniques for handling the comparison of several means with unknown and possibly unequal variances. There is, however, no excuse whatever for our not making use of Welch's results for the comparison of two means. This is none other than the Behrens-Fisher problem (famous or infamous, as your sympathies lie) over which so much controversy has been waged. Behrens' work appeared in 1929; Fisher's fiducial inference argument dating from 1930 led to corroboration of Behrens' work in 1941. You may not accept the fiducial test as satisfactory; there are many quite able statisticians who do not. Kendall leaves his reader with "a choice of several attitudes toward the foundations of the fiducial argument: (a) he can accept the argument as involving a new postulate of inference; he can regard it as sanctioned by" essentially Fisher's approach; "or (c) he can, so far as estimates based on a single parameter are concerned, console himself with the thought that results of the process are the same as those given by the theory of confidence intervals." He points out that an α-level Behrens-Fisher fiducial test of the difference between two means does not insure that the user will be correct in (1-α) percent of cases; rather he recommends Welch's procedure to insure this kind of protection. Welch observes that . . . "although Fisher's approach has been very much criticized by a number of writers, starting with Bartlett(1936), the critics have not wished to throw doubt on the whole body of results that Fisher includes under the heading of fiducial inference." See Welch 1947, p.34. Welch's work dates from 1938 and the results were tabled by Aspin in 1949. It is unfortunate that these tables contain the critical values for only the 1 percent and 5 percent one tailed test corresponding to the 2 percent and 10 percent values for the two tailed test, but in my opinion these tables should be more widely disseminated and used. The procedure is as easily applied as the more usual t-test and in most cases much more appropriate. I want to illustrate its use with a couple of examples and recommend that you use it. This was, principally, what I had in mind when I suggested the title of this paper as a topic for discussion. It is not appropriate here to draw further comparisons between the Behrens-Fisher and Aspin-Welch procedures; however, I hope to include as an appendix an elementary treatment of such a comparison. I do want to include in the body of the paper, however, what we know about the more general case of more than two means.

A problem that has arisen on several occasions in my own office and that embodies almost all of the difficulties associated with the problem of unequal variances is the following. Cooperative fluid milk handlers pay their patrons monthly in proportion to the amount of butterfat delivered. The amount of butterfat is computed on the combined basis of total pounds of milk delivered and a Babcock test of the percent butterfat. In an effort to reduce the amount of testing, the procedure has been fairly frequently adopted of blending daily samples into composite samples, adding an inhibitor and storing under refrigeration for 7, 10 or even 15 days before making the Babcock test. It has been fairly well established that through storage the composites generally lead to estimates of monthly butterfat deliveries that are biased downward. The amount of the bias is usually slight but increases in general with increased storage time, varies somewhat directly with the percent

butterfat, varies from producer to producer and from season to season (with temperature) for the same producer. Add to this the fact that the percent butterfat itself varies more or less periodically pretty much in phase with the seasons.

The kind of data one is usually confronted with is similar to that in Table 4 expanded somewhat to include more months and more producers but rarely expanded in the direction of more measurements per day. It is quite natural to expect that the variances of monthly means for the four sampling methods are far from equal. We will see in a moment that in our example the variance among daily tests within any month is quite different from that among 7-day composite tests. Keeping in mind our comments on the L.S.D. test and reminding you again of the changes from producer to producer and from month to month in both mean and variances, I would like to throw out for your suggestions the question of how best to analyze such data so as to ascertain whether the compositing procedure does in fact lead to biased results and if so whether the amount of bias is related to the length of the compositing period. I am quite serious in soliciting your remarks as to proper techniques for handling such data.

For our present purposes, we will be interested less in the practical question of bias and how to test for it than in a simple application of the test for the difference between two means when the variances cannot be assumed to be equal. Table 4 shows the results of butterfat tests for one producer for the months of August and September of 1951 and for June of 1952. Let us first test whether the difference between the mean for the two-month Fall 1951 season and the mean for June of the following year is significantly positive. The relevant statistics are:

$m_1 = 4.1836 \qquad m_2 = 3.8700$

$s_1^2 = .24073 \qquad s_2^2 = .02148$

$n_1 = 60 \qquad n_2 = 29$

where the m's represent the sample means, the s^2's represent the sample estimates of variance and the n's are the respective degrees of freedom.

If we could assume equal variances we would compute

$$t = \frac{m_1 - m_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\left(\frac{1}{n_1+1} + \frac{1}{n_2+1}\right)}} = 3.42$$

and compare with tabular values in a t-table at $n_1 + n_2 = 89$ d.f. or with tabular values in a table of areas under the Normal curve. Alternatively, one might compute

$$t' = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1+1} + \frac{s_2^2}{n_2+1}}} = 4.55$$

There is some question about what d.f. to use in testing t'. Some people use $n_1 + n_2$; others use the minimum of n_1 and n_2; still others always refer t' to Normal tables perhaps even for small sample sizes. Welch has proposed as the effective d.f., f, for t': $\frac{1}{f} = \frac{c^2}{n_1} + \frac{1-c^2}{n_2}$ where $o \leq c \leq 1$ so that f always lies between the smaller of the n_i and their sum. C will be defined in a moment. See Welch (1949) p. 296.

It is quite well known that neither t nor t' is distributed as Students' t if $\sigma_1 = \sigma_2$; if $\sigma_1 = \sigma_2$, t is distributed as Students' t while t' is not unless in addition $n_1 = n_2$ in which case $t = t'$. It may be because of the equality in the last case that the degrees of freedom ascribed to t' are sometimes taken to be $n_1 + n_2$; clearly the use of the minimum of n_1 and n_2 is an attempt to provide a conservative test. If it is known that $\sigma_1 = \sigma_2$, then t is surely the statistic to use. If, however σ_1 and σ_2 differ, then t may give very misleading results and it will be safer to use t'. Figure 2 is a reproduction of Welch's Figure 1 1/ from the reference cited below and shows that the unknown ratio of variances ω need not differ much from unity before v (our t') becomes less biased than u (our t) at least for the case illustrated of $n_1 = 4$ and $n_2 = 14$. Welch's test, which is to be recommended, employs the statistic t' but requires that the significance of the statistic be judged by comparison with the Aspin-Welch tables as provided in Appendix tables 2 and 3. The only calculation required to enter the table is

$$s_2^2/n_2+1 \Big/ (s_1^2/n_1+1 + s_2^2/n_2+1) = .15.$$

Again you will hasten to point out to me that both t and t' are well beyond any of the tabular values for either the 5 percent or 1 percent Student t or Aspin-Welch t', respectively. Both exceed the one tailed 1 percent Normal deviate. However, I would like to counter with the reminder that t is just under 3-1/2 while t' just exceeds 4-1/2; the area under the Normal curve beyond the standard Normal deviates 3-1/2 and 4-1/2, though both quite small are relatively quite different in magnitude.

As a second example, let's perform the same test for the 7-day composites. Here the relevant statistics are

$m_1 = 4.175$ $\qquad$ $m_2 = 3.850$

$s_1^2 = .07643$ $\qquad$ $s_2^2 = .00333$

$n_1 = 7$ $\qquad$ $n_2 = 3$

from which we find that $t = 2.28$ and $t' = 3.19$. The first of these, when compared with the one tailed tabular Student t with 10 d.f., falls far short of significance at the 1 percent level; in fact the probability of a Student t as large or larger than 2.28 exceeds .02. On the other hand $t' = 3.19$ at

---

1/ Welch, B. L. Biometrika 1938. p. 355.

10 d.f. is significant at the 1 percent level; the probability of a Student t as large or larger than 3.19 is less than .005. As in the previous example, there is little need for computing the ratio

$$\frac{s_1^2/n_1+1}{s_1^2/n_1+1 + s_2^2/n_2+1}$$

because t′ exceeds every entry in the Aspin-Welch table of one-tailed 1 percent level test tabular values.

In passing, we might note that examination of the differences between daily mean tests and the corresponding 7-day composite tests does not, in this particular instance, lead one to suspect unequal variances; furthermore, the amount of bias, though positive, is not sufficiently large to establish its significance on so few degrees of freedom.

In order to illustrate further the use of the Aspin-Welch tables suppose we test for the difference between means in the following example:

$$m_1 = 60.3 \qquad s_1^2 = 51 \qquad n_1 = 9$$

$$m_2 = 47.1 \qquad s_2^2 = 340 \qquad n_2 = 6$$

computing $$\frac{s_1^2/n_1+1}{s_1^2/n_1+1 + s_2^2/n_2+1} = \frac{5.1}{53.67} = .095$$

and $$t' = \frac{m_1 - m_2}{\sqrt{53.67}} = \frac{13.2}{7.326} = 1.80$$

Entering Table 3 with $\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2} = .095 \doteq .1$, $f_2 = 6$, $f_1 = 9$ we find that t′ = v needs to be 1.90 to be significant at the 5 percent level. The one tailed Student t required for 15 d.f. and at the 5 percent level of significance is only 1.753. Here then, interpreting t′ as a, Student t leads falsely to the conclusion that the means are significant at the 5 percent level. If instead of t′ we had computed

$$t = \frac{m_1 - m_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_2 + n_2}\left(\frac{1}{n_1+1} + \frac{1}{n_2+1}\right)}} = \frac{13.2}{6.361} = 2.08 ,$$

again we would have been led to the erroneous conclusion that the means differ significantly. It might be worth noting in passing that Welch's proposed effective degrees of freedom are f = 6 for which the 5 percent one tailed, critical value is 1.94, which agrees fairly closely with the value 1.90 from Table 3.

The use of the Aspin-Welch tables for estimating confidence intervals is precisely the same as the use of the Student-t tables for the same purpose.

Thus, in the example just completed, we would have: (1) in keeping with our use of the one tailed, a lower confidence limit of zero for the difference between the two means and (2) because

$$P\left\{-\infty < t' < 1.90\right\} = .95$$

$$\text{i.e. } P\left\{-\infty < \frac{m_1 - m_2}{7.326} < 1.90\right\} = .95,$$

then the upper 95 percent confidence limit for the difference is 7.326 x 1.90 = 13.9.

In closing the discussion of the test of the significance of two means, I should point out to those interested in the mathematical aspects of the problem the elegant treatment by Hsu in the reference cited.

A few highlights from Hsu's paper are:

Defining $\delta = \mu_1 - \mu_2$, $\Theta = \sigma_1^2/\sigma_2^2$, $\sigma^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$, where $N_1 = n_1+1$, $N_2 = n_2+1$

$$\lambda = \frac{\delta^2}{2\sigma^2}$$

$$u = (m_1 - m_2)^2 / (A_1\Sigma_1 + A_2\Sigma_2)$$

$$B_j\sigma^2 = A_j\,\sigma_j^2$$

Hsu gets the distribution of u in general.

Special cases are

(1) u_1 = our t^2 comes from taking

$$A_1 = A_2 = N/(N-2)\,N_1N_2 \qquad N = N_1+N_2$$

$$\frac{B_1}{\Theta} = B_2 = N/(N-2)\,(N_1+N_2\Theta)$$

$$B_1/B_2 = \Theta$$

(2) u_2 = our t'^2 comes from taking

$$A_j = 1/N_j\,(N_j-1)$$

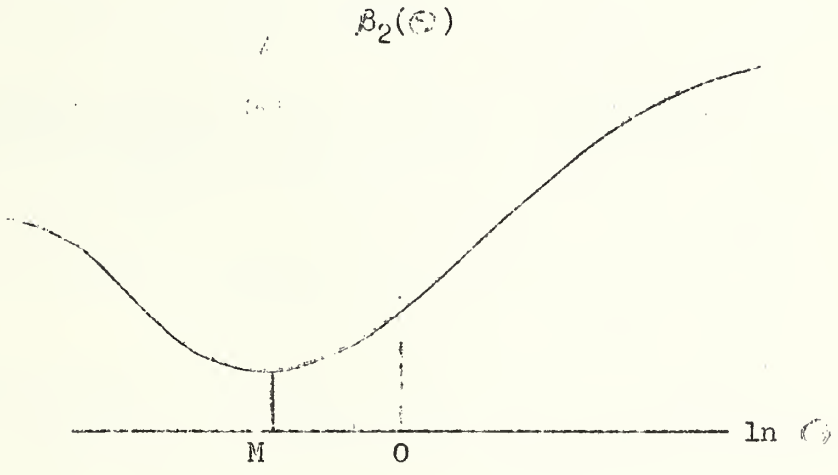
In particular he shows:

(1) For any N_1, N_2 and fixed Θ the <u>power</u> of u, $\delta(\lambda,\Theta)$, is increasing with λ.

(2) $\beta(o, \Theta) = \beta(\Theta)$ increases with Θ for $\Theta > 1$.

(3) $\beta(\lambda, \Theta, N_1, N_2) = \beta(\lambda, \Theta^{-1}, N_2, N_1)$

(4) For $\lambda = o$ $\beta(\lambda, \Theta) = \beta(\Theta)$ = Size of the test, in contrast to the power. (a) For $u = u_2$ = our t^2, and for $N_1 < N_2$, $\beta_2(\Theta)$ takes the form



but the infinite branch of the curve, as $\log \Theta \longrightarrow -\infty$, may not rise as high as $\beta_2(1)$. $M = \ln \dfrac{N_1(N_1-1)}{N_2(N_2-1)}$. If $N_1 > N_2$ the curve is simply reflected in the vertical axis.

(b) The form of $\beta_1(\Theta)$ is similar to that of $\beta_2(\Theta)$ except that for $N_2 - N_1 > 2$ the curve declines as $\ln \Theta \longrightarrow -\infty$.

(c) For $N_1 = N_2$, $u_1 = u_2$ or $t = t'$; so $\beta_1(\Theta) = \beta_2(\Theta)$. The point M moves to zero, and, of course, the curve is symmetrical.

His conclusions are as follows:

(A) If the hypothesis to be tested is H_1: $\lambda = 0, \Theta = 1$, (not the hypothesis we have considered in this paper, namely $\lambda = 0$) sampling from certain populations u_1, or t, will reject H_1 less frequently when it is false than when it is true, which is clearly a most unsatisfactory result. u_2, or t', is less seriously biased and over a considerably more restricted domain

of points $(\lambda, \quad)$ but is likely to be less sensitive to variations of . For $N_1 = N_2$ (i.e. $u_1 = u_2$) the common test is unbiased but quite insensitive to variations in . Hsu would rule out both u_1 and u_2 as inadequate for testing H_1.

(B) If the hypothesis to be tested is H_2: $\lambda = 0$ (the test discussed in this paper) both u_1 and u_2 have $\beta(\lambda,)$ a minimum at $\lambda = 0$ (i.e., unbiased) but do not have distributions that are independent of ; so it is impossible to control the type I error. "The test u_2 has the advantage over u_1 because it is likely to be more insensitive to the variation of . . . . If $N_1 = N_2 = n$ and if n is not very small, then $u_1 (= u_2)$ is so indifferent to the variation of that we may safely use it to test H_2 just as if were known to be unity. In fact, if, say, n = 10, then, in using the u_1-test for H_2 and taking the significance level as the 5 percent level of t^2, we can say that (i) the test is unbiased, (ii) whenever H_2 is true, the chance of rejecting it is between 0.05 and 0.065, and (iii) if = 1, then there is no other unbiased test, adjusted to give the same chance 0.05 of the first kind error, which is more able to detect a departure of λ from 0."

The problem of testing the equality of several means is, of course, much more difficult than that of testing the equality of two means. If the problem of two means has, itself, proved somewhat intractable and satisfactory methods for handling it only relatively recently made available, it is natural to expect that very little is known about how to handle the more difficult problem. There are, in fact, only a few papers that provide the applied statistician any constructive suggestions.

We noted a little earlier that the quantity $\sum_{i=1}^{k} \frac{n_i}{\sigma_1^2} (m_i - \bar{m})^2$ is distributed as χ^2 with k-1 d.f. if the k unknown population means are all equal without any assumption about equality of the σ_i^2. This provides the basis for

a convenient large sample approximate test by replacing the σ_i^2 by their sample estimates s_i^2. (in the coefficients and in $\bar{m}$) and still using the χ^2 tables to judge the significance; the resulting statistic affords a further challenge to obtain its small sample distribution. This is not at all easy to obtain. G. S. James, in a logical attempt to studentize the problem, replaces the probability $P\left[\sum \frac{n_i}{\sigma_i^2}(m_i - \bar{m})^2 > \chi^2(\alpha)\right] = \alpha$ by the corresponding probability $P\left[\sum \frac{n_i}{s_i^2}(m_i - \bar{m})^2 > f(s_i^2, \alpha)\right]$ and obtains for the required function f a power series in χ^2 s. Taking only the first corrective terms (terms through those of order -1 in the n_i) he recommends, as an improvement over the χ^2 approximation test, the use of

$$\chi^2\left[1 + \frac{3\chi^2 + (k+1)}{2(k^2 - 1)} \sum \frac{1}{n_i}\left(1 - \frac{w_i}{w}\right)^2\right]$$

as the critical value of the statistic $\sum w_i(m_i - \bar{m})$, where $w_i = \frac{n_i}{s_i^2}$ and $w = \sum \frac{n_i}{s_i^2}$. (He also gives the term of order -2 in the n_i but notes because of complexity that it will likely prove of not too much practical utility.)

Welch, as in his earlier work on the problem of two means, approximates the distribution of James' statistic by comparing its moment generating function with that of the familiar F distribution. He defines a new statistic

$$v^2 = \frac{\sum w_i(m_i - \bar{m})^2 \,/\, k - 1}{\left[1 + \frac{2(k-2)}{k^2-1} \sum \frac{1}{n_i}\left(1 - \frac{w_i}{w}\right)^2\right]}$$

the significance of which is to be judged by entering the F table with degrees of freedom k - 1 and $\left[\frac{3}{k^2-1} \sum \frac{1}{n_i}\left(1 - \frac{w_i}{w}\right)^2\right]^{-1}$ and shows that to order -1 in the n_i this procedure is equivalent to James' procedure. "To higher orders, of course, the two procedures differ. There are obvious points which could be mentioned in favor of either method, but no extensive numerical work has been done to compare them." Example 1 appended is an excerpt from Welch's paper.

Related to the problem of how to test for equality among several means when variances may not be equal is the question of how the inequality of variances affects the analysis of variance test of equality of means. The

papers by Box, Quensel, and David and Johnson cited in the references are concerned more with this aspect of the problem than with how properly to test for equality.

An interesting example of the Behrens-Fisher problem extended to multivariate dimensions was considered recently by Fraser 20/ in comparing q of p + q regression coefficients in one normal population with the corresponding q coefficients in another population where the respective residual variances could not be assumed to be equal.

We have spoken here today only of testing collectively for equality of means. Once the question of equality has been decided in the negative there immediately arises the question of how to separate the disparate means. In case variances cannot be assumed to be equal, we know little or nothing about the solution to this problem; if variances are equal there are a choice of recently developed methods to apply. A discussion of the relative merits of these methods is to be the subject of one of our seminars in the very near future.

Selected References on the Test of Equality of Means
When Variances May be Unequal

(1) "Behren's, W. V. (1929) "Ein Beitrag zur Fehlerberechnung bei weinige Beobachtungen." Landw. Jb. 68, 807-37.

(2) Fisher, R. A. (1935) The Fiducial Argument in Statistical Inference, Ann. Eug., Lond. 6, 391-8.

(3) Fisher, R. A. (1937) On a point raised by M. S. Bartlett on Fiducial Probability, Ann. Eug., Lond. 7, 370-5.

(4) Fisher, R. A. (1939) Samples with possibly unequal variances, Ann. Eug. Lond. 9, 174-80.

(5) Welch, B. L. "The significance of the difference between two means when the population variances are unequal," Biometrika, Vol. 29 (1937,) pp. 350-362.

(6) Bartlett, M. S. (1936) The Information Available in Small Samples. Proc. Camb. Phil. Soc. 32, 560-6.

(7) Sukhatme, P. V. (1938) On Fisher and Behren's test of significance for the difference in means of 2 normal samples, Sankhya 4, 39-48.

(8) Wilks, S. S. On the Problem of two samples from Normal Populations with Unequal Variances, A.M.S. 11 475-76. (1940).

(9) Hsu, P. L. "Contributions to the theory of 'Students' t-test as applied to the problem of two samples," Statistical Research Memoirs, Vol. 2 (1938), pp. 1-24.

(10) Gronow, D.G.C. Test for the Significance of the Difference between Means of two Normal Populations having Unequal Variances, Biometrika 38 (1951).

(11) Quensel, C. E. "The validity of the z-criterion when the variates are taken from different normal populations," Skand. Aktuarietids., Vol. 30 (1947), pp. 44-55.

(12) David, F. N. and Johnson, N. L. "The effect of non-normality on the power function of the F-test in the analysis of variance," Biometrika, Vol. 38 (1951), pp. 43-57.

(13) David, F. N. and Johnson, N. L. "A method of investigating the effect of non-normality and heterogeneity of variance on tests of the general linear hypothesis," Ann. Math. Stat., Vol. 22 (1951), pp. 382-392.

(14) David, F. N. and Johnson, N. L. "The sensitivity of analysis of variance tests with respect to random between groups variation," Trabajos de Estadistica, Vol. 2 (1951) pp. 179-188.

Selected references (Contd)

(15) James, G. S. The Comparison of Several Groups of Observations when the Ratios of the Population Variances are Unknown. Biometrika 38, pts 3 and 4 (1951) pp. 324-329.

(16) Welch, B. L. On the Comparison of Several Mean Values: An Alternative Approach. Biometrika 38, Pts. 3 and 4 (1951) pp. 330-336.

(17) Neyman, J. Fiducial Argument and the Theory of Confidence Intervals, Biometrika 32 (1941) p. 128.

(18) Scheffe, Henry. On Solutions of the Behrens-Fisher Problem, Based on the t- Distribution, Annals of Math. Stat. XIV (1943) pp 35-44.

(19) Chand, Uttam. Distributions Related to Comparison of Two Means and Two Regression Coefficients, Annals of Math. Stat. XXI (1950) pp 507-522.

(20) Fraser, D. A. S. The Behrens-Fisher Problem for Regression Coefficients, Annals of Math. Stat. XXIV (1953) pp 390-402.

(21) Gronow, D.G.C. Non Normality in two sample t-tests, Biometrika 40(1953) pp 222-225.

(22) Horsnell, G. The Effect of Unequal Variances on the F-test for the Homogeneity of Group Means, Biometrika 40(1953) pp 128-136.

Figure 1. Comparison of Mean and Standard Deviation

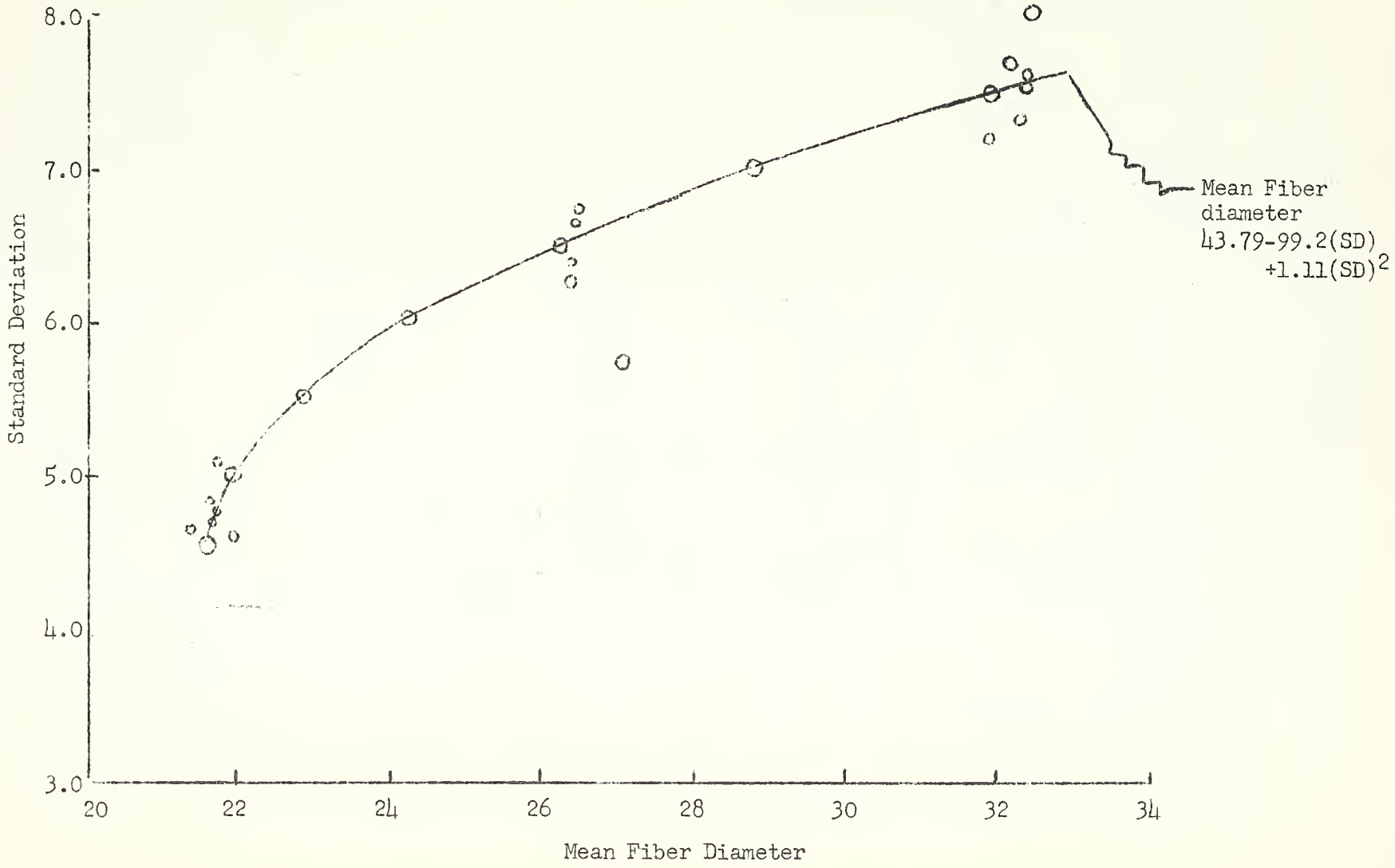


Figure 2. Comparison of Usual Tests

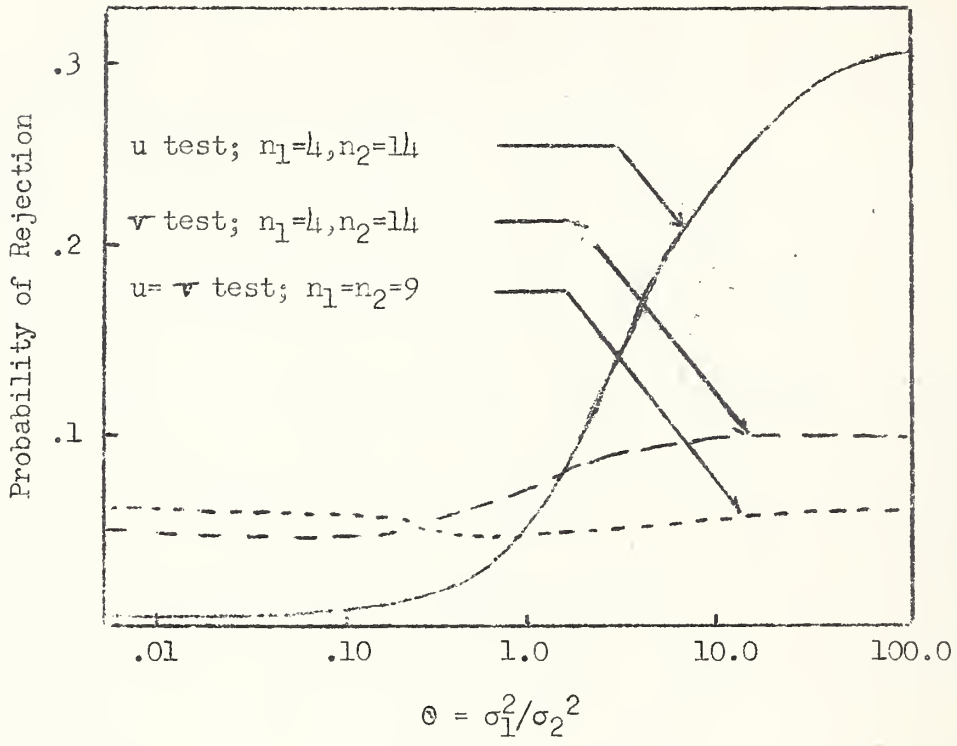


Figure 3. Comparison of Usual Tests with Aspin-Welch Test n_1= 4; n_2 = 14

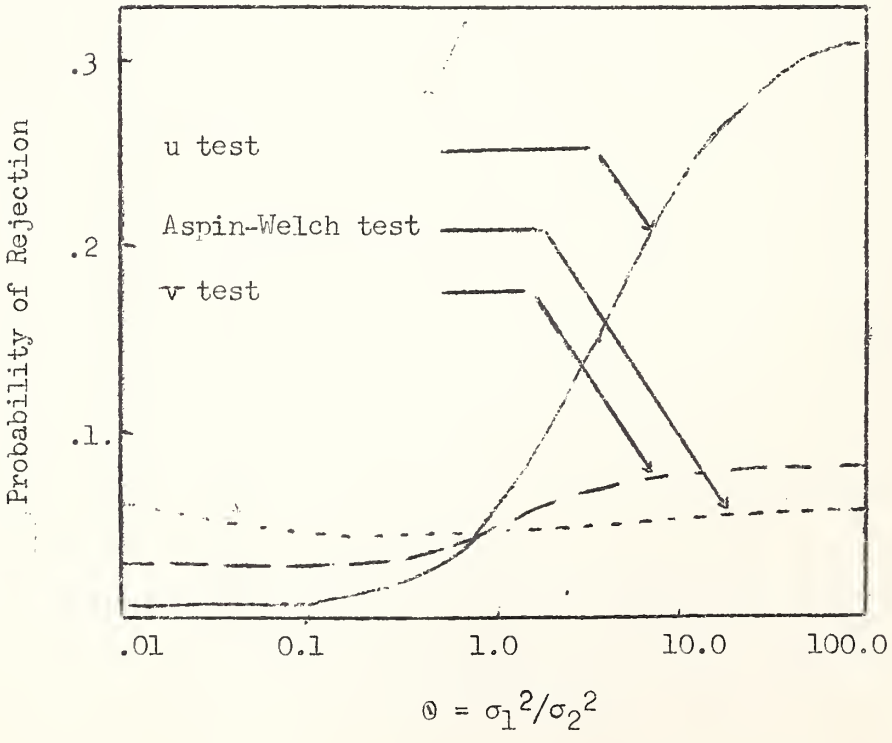


Table 1. Average fiber diameter, standard deviation, and coefficient of variation for scoured core and card sliver samples.

Grade of lot	Number of bales cored	Number of cores drawn per bale	Procedure for core sampling	Plan of sub-sampling	Number of fibers measured	Average fiber diameter	Standard deviation	Coefficient of variation
						Microns	Microns	Percent
64s staple	10	10	Composite 3/8"	I	600	21.78	4.80	22.04
				II	600	21.40	4.63	21.64
	10	5	Individual 3/8"		5,000	21.76	5.05	23.21
	10	10	Composite 1-1/4"	I	600	21.70	4.85	22.35
				II	600	21.70	4.71	21.70
	Card sliver				600	21.90	4.63	21.14
56/58s staple	10	8	Composite 3/8"	I	1,000	26.35	6.53	24.78
				II	1,000	26.58	6.80	25.58
	10	5	Individual 3/8"		5,000	26.61	6.70	25.18
	10	8	Composite 1-1/4"	I	1,000	26.98	5.80	21.50
				II	1,000	26.50	6.48	24.45
	Card sliver				1,000	26.60	6.33	23.80
48/46s staple	2	8	Composite 3/8"	I	1,000	31.80	7.25	22.80
				II	1,000	32.38	7.63	23.56
	2	5	Individual 3/8"		1,000	32.40	7.55	23.30
	2	8	Composite 1-1/4"	I	1,000	32.30	8.05	24.92
				II	1,000	32.25	7.38	22.88
	Card sliver				1,000	32.12	7.71	24.00

Table 2. Value of $v = \frac{(y - \eta)}{\sqrt{(\lambda_1 s_1^2 + \lambda_2 s_2^2)}}$ exceeded with probability $\epsilon = 0.01$*

[or of / v / exceeded with probability $2\epsilon = 0.02$]

	$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
f_2=10	f_1= 10	2.76	2.70	2.63	2.56	2.51	2.50	2.51	2.56	2.63	2.70	2.76
	12	2.76	2.70	2.63	2.56	2.51	2.49	2.49	2.52	2.57	2.62	2.68
	15	2.76	2.70	2.63	2.56	2.51	2.48	2.47	2.48	2.52	2.56	2.60
	20	2.76	2.70	2.63	2.56	2.51	2.47	2.45	2.45	2.47	2.49	2.53
	30	2.76	2.70	2.63	2.56	2.50	2.46	2.43	2.42	2.42	2.44	2.46
	∞	2.76	2.70	2.63	2.56	2.50	2.44	2.40	2.36	2.34	2.33	2.33
f_2=12	f_1= 10	2.68	2.62	2.57	2.52	2.49	2.49	2.51	2.56	2.63	2.70	2.76
	12	2.68	2.62	2.57	2.52	2.48	2.47	2.48	2.52	2.57	2.62	2.68
	15	2.68	2.62	2.57	2.52	2.48	2.46	2.46	2.48	2.52	2.56	2.60
	20	2.68	2.62	2.57	2.52	2.48	2.45	2.44	2.45	2.47	2.49	2.53
	30	2.68	2.62	2.57	2.52	2.47	2.44	2.42	2.41	2.42	2.44	2.46
	∞	2.68	2.62	2.57	2.51	2.46	2.42	2.38	2.36	2.34	2.33	2.33
f_2=15	f_1= 10	2.60	2.56	2.52	2.48	2.47	2.48	2.51	2.56	2.63	2.70	2.76
	12	2.60	2.56	2.52	2.48	2.46	2.46	2.48	2.52	2.57	2.62	2.68
	15	2.60	2.56	2.51	2.48	2.45	2.45	2.45	2.48	2.51	2.56	2.60
	20	2.60	2.56	2.51	2.48	2.45	2.43	2.43	2.44	2.46	2.49	2.53
	30	2.60	2.56	2.51	2.47	2.44	2.42	2.41	2.41	2.42	2.44	2.46
	∞	2.60	2.56	2.51	2.47	2.43	2.40	2.37	2.35	2.34	2.33	2.33
f_2=20	f_1= 10	2.53	2.49	2.47	2.45	2.45	2.47	2.51	2.56	2.63	2.70	2.76
	12	2.53	2.49	2.47	2.45	2.44	2.45	2.48	2.52	2.57	2.62	2.68
	15	2.53	2.49	2.46	2.44	2.43	2.43	2.45	2.48	2.51	2.56	2.60
	20	2.53	2.49	2.46	2.44	2.42	2.42	2.42	2.44	2.46	2.49	2.53
	30	2.53	2.49	2.46	2.44	2.42	2.40	2.40	2.40	2.42	2.43	2.46
	∞	2.53	2.49	2.46	2.43	2.40	2.38	2.36	2.34	2.33	2.33	2.33
f_2=30	f_1= 10	2.46	2.44	2.42	2.42	2.43	2.46	2.50	2.56	2.63	2.70	2.76
	12	2.46	2.44	2.42	2.41	2.42	2.44	2.47	2.52	2.57	2.62	2.68
	15	2.46	2.44	2.42	2.41	2.41	2.42	2.44	2.47	2.51	2.56	2.60
	20	2.46	2.43	2.42	2.40	2.40	2.40	2.42	2.44	2.46	2.49	2.53
	30	2.46	2.43	2.42	2.40	2.39	2.39	2.39	2.40	2.42	2.43	2.46
	∞	2.46	2.43	2.41	2.39	2.37	2.36	2.35	2.34	2.33	2.33	2.33
f_2=∞	f_1= 10	2.33	2.33	2.34	2.36	2.40	2.44	2.50	2.56	2.63	2.70	2.76
	12	2.33	2.33	2.34	2.36	2.38	2.42	2.46	2.51	2.57	2.62	2.68
	15	2.33	2.33	2.34	2.35	2.37	2.40	2.43	2.47	2.51	2.56	2.60
	20	2.33	2.33	2.33	2.34	2.36	2.38	2.40	2.43	2.46	2.49	2.53
	30	2.33	2.33	2.33	2.34	2.35	2.36	2.37	2.39	2.41	2.43	2.46
	∞	2.33	2.33	2.33	2.33	2.33	2.33	2.33	2.33	2.33	2.33	2.33

* y is normally distributed about η with variance $(\lambda_1\sigma_1^2 + \lambda_2\sigma_2^2)$, and s_1^2 and s_2^2 are independent estimates of σ_1^2 and σ_2^2, based on f_1 and f_2 degrees of freedom, respectively. λ_1 and λ_2 are known constants.

In the problem of comparing the means of samples taken from two normal populations, put $y = (\bar{x}_1 - \bar{x}_2)$, $f_1 = (n_1 - 1)$, $f_2 = (n_2 - 1)$, $\lambda_1 = 1/n_1$ and $\lambda_2 = 1/n_2$, where n_1 and n_2 are the sample sizes.

Table 3. Value of $v = \frac{(y - \eta)}{\sqrt{(\lambda_1 s_1^2 + \lambda_2 s_2^2)}}$ exceeded with probability $\epsilon = 0.05$*

[or of / v / exceeded with probability $2\epsilon = 0.10$]

	$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$f_2 = 6$	$f_1 = 6$	1.94	1.90	1.85	1.80	1.76	1.74	1.76	1.80	1.85	1.90	1.94
	8	1.94	1.90	1.85	1.80	1.76	1.73	1.74	1.76	1.79	1.82	1.86
	10	1.94	1.90	1.85	1.80	1.76	1.73	1.73	1.74	1.76	1.78	1.81
	15	1.94	1.90	1.85	1.80	1.76	1.73	1.71	1.71	1.72	1.73	1.75
	20	1.94	1.90	1.85	1.80	1.76	1.73	1.71	1.70	1.70	1.71	1.72
	∞	1.94	1.90	1.85	1.80	1.76	1.72	1.69	1.67	1.66	1.65	1.64
$f_2 = 8$	$f_1 = 6$	1.86	1.82	1.79	1.76	1.74	1.73	1.76	1.80	1.85	1.90	1.94
	8	1.86	1.82	1.79	1.76	1.73	1.73	1.73	1.76	1.79	1.82	1.86
	10	1.86	1.82	1.79	1.76	1.73	1.72	1.72	1.74	1.76	1.78	1.81
	15	1.86	1.82	1.79	1.76	1.73	1.71	1.71	1.71	1.72	1.73	1.75
	20	1.86	1.82	1.79	1.76	1.73	1.71	1.70	1.70	1.70	1.71	1.72
	∞	1.86	1.82	1.79	1.75	1.72	1.70	1.68	1.66	1.65	1.65	1.64
$f_2 = 10$	$f_1 = 6$	1.81	1.78	1.76	1.74	1.73	1.73	1.76	1.80	1.85	1.90	1.94
	8	1.81	1.78	1.76	1.74	1.72	1.72	1.73	1.76	1.79	1.82	1.86
	10	1.81	1.78	1.76	1.73	1.72	1.71	1.72	1.73	1.76	1.78	1.81
	15	1.81	1.78	1.76	1.73	1.72	1.70	1.70	1.71	1.72	1.73	1.75
	20	1.81	1.78	1.76	1.73	1.71	1.70	1.69	1.69	1.70	1.71	1.72
	∞	1.81	1.78	1.76	1.73	1.71	1.69	1.67	1.66	1.65	1.65	1.64
$f_2 = 15$	$f_1 = 6$	1.75	1.73	1.72	1.71	1.71	1.73	1.76	1.80	1.85	1.90	1.94
	8	1.75	1.73	1.72	1.71	1.71	1.71	1.73	1.76	1.79	1.82	1.86
	10	1.75	1.73	1.72	1.71	1.70	1.70	1.72	1.73	1.76	1.78	1.81
	15	1.75	1.73	1.72	1.70	1.70	1.69	1.70	1.70	1.72	1.73	1.75
	20	1.75	1.73	1.72	1.70	1.69	1.69	1.69	1.69	1.70	1.71	1.72
	∞	1.75	1.73	1.72	1.70	1.68	1.67	1.66	1.65	1.65	1.65	1.64
$f_2 = 20$	$f_1 = 6$	1.72	1.71	1.70	1.70	1.71	1.73	1.76	1.80	1.85	1.90	1.94
	8	1.72	1.71	1.70	1.70	1.70	1.71	1.73	1.76	1.79	1.82	1.86
	10	1.72	1.71	1.70	1.69	1.69	1.70	1.71	1.73	1.76	1.78	1.81
	15	1.72	1.71	1.70	1.69	1.69	1.69	1.69	1.70	1.72	1.73	1.75
	20	1.72	1.71	1.70	1.69	1.68	1.68	1.68	1.69	1.70	1.71	1.72
	∞	1.72	1.71	1.70	1.68	1.67	1.66	1.66	1.65	1.65	1.65	1.64
$f_2 = \infty$	$f_1 = 6$	1.64	1.65	1.66	1.67	1.69	1.72	1.76	1.80	1.85	1.90	1.94
	8	1.64	1.65	1.65	1.66	1.68	1.70	1.72	1.75	1.79	1.82	1.86
	10	1.64	1.65	1.65	1.66	1.67	1.69	1.71	1.73	1.76	1.78	1.81
	15	1.64	1.65	1.65	1.65	1.66	1.67	1.68	1.70	1.72	1.73	1.75
	20	1.64	1.65	1.65	1.65	1.66	1.66	1.67	1.68	1.70	1.71	1.72
	∞	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.64	1.64

*y is normally distributed about η with variance $(\lambda_1\sigma_1^2 + \lambda_2\sigma_2^2)$, and s_1^2 and s_2^2 are independent estimates of σ_1^2 and σ_2^2, based on f_1 and f_2 degrees of freedom, respectively. λ_1 and λ_2 are known constants.

In the problem of comparing the means of samples taken from two normal populations, put $y = (\bar{x}_1 - \bar{x}_2)$, $f_1 = (n_1 - 1)$, $f_2 = (n_2 - 1)$, $\lambda_1 = 1/n_1$ and $\lambda_2 = 1/n_2$, where n_1 and n_2 are the sample sizes.

Table 4. Comparison of butterfat tests for a selected producer

Date	Daily test	Mean daily test	7 day composite test	Difference between (3)&(4)	10 day composite test	15 day composite test	Date	Daily test	Mean daily test	7 day composite test	Difference between (3)&(4)	10 day composite test	15-day composite test
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Aug.1951							Sept.1951						
1	4.2						1	3.4					
2	3.3						2	4.2					
3	3.6						3	4.0					
4	4.6						4	4.2					
5	4.6						5	3.9					
6	3.7						6	4.1					
7	4.2	4.029	4.0	.029			7	4.6	4.057	4.1	-.043		
8	3.7						8	4.4					
9	4.3						9	3.7					
10	4.3				4.0		10	4.6				4.1	
11	5.2						11	4.2					
12	5.5						12	3.9					
13	3.8						13	4.3					
14	4.0						14	3.8					
15	4.0	4.350	4.4	-.050		4.1	15	4.9	4.225	4.2	.025		4.2
16	3.3						16	4.0					
17	3.9						17	4.5					
18	4.2						18	3.9					
19	3.8						19	4.1					
20	3.8				4.2		20	4.5				4.3	
21	4.4						21	4.3					
22	4.4						22	4.1	4.200	4.2	.000		
23	4.1	3.988	4.0	-.012			23	4.9					
24	4.1						24	4.8					
25	3.9						25	4.3					
26	4.5						26	4.2					
27	3.4						27	5.4					
28	3.9						28	4.6					
29	3.3						29	4.4					
30	4.3						30	5.2	4.725	4.7	.025	4.6	4.5
31	3.5	3.862	3.8	.062	4.0	4.0	Mo.Mean	4.313	4.302	4.300	.002	4.333	4.350
Mo.Mean	4.058	4.057	4.050	.007	4.067	4.050	Overall Mean	4.184	4.180	4.175	.004	4.200	4.200

(Continued) Table 4. Comparison of butterfat tests for a selected producer

Date	Daily test	Mean daily test	7 day composite test	Difference between (3)&(4)	10 day composite test	15 day composite test
(1)	(2)	(3)	(4)	(5)	(6)	(7)
June 1952						
1	3.8					
2	3.7					
3	3.7					
4	3.9					
5	3.9					
6	3.8					
7	3.9	3.814	3.8	.014		
8	4.1					
9	3.7					
10	3.8				3.3	
11	4.1					
12	3.9					
13	3.9					
14	3.6					
15	3.9	3.875	3.8	.075		3.2
16	3.8					
17	3.8					
18	4.0					
19	4.0					
20	4.0					
21	3.7				3.3	
22	4.0	3.900	3.9			
23	3.8					
24	3.6					
25	4.0					
26	3.7					
27	4.1					
28	3.9					
29	4.1					
30	3.9	3.888	3.9	-.012	3.3	3.3
Mean	3.870	3.869	3.850	.019	3.300	3.250

Example 1

The data in Table 1 relate to an experiment in which three treatments are being compared, a single characteristic x being measured. As mentioned in paragraph 1 above, the y_t are now the means $\bar{x}_t$ for the three treatments. The s_t^2 are the individual 'within treatment' variances, estimated on degrees of freedom f_t, equal to one less than the number of replicates in each case. Also $\lambda_t = 1/n_t$. Hence, $w_t = n_t/s_t^2$.

Table 1

Treatment (t)	Number of individuals (n_t)	Treatment mean ($\bar{x}_t$)	Observed variance (s_t^2)	Estimated variance of mean (s_t^2/n_t)	W_t (n_t/s_t^2)
1	20	27.8	60.1	3.00	0.333
2	10	24.1	6.3	0.63	1.587
3	10	22.2	15.4	1.54	0.649
					2.569

If the true means μ_t are all equal, the best estimate of the common μ will be given by $\hat{x} = (\sum w_t x_t)/(\sum w_t) = 24.10$. The details required in this calculation and further quantities needed later are shown in Table 2. An arbitrary origin, x = 25, is being used.

Table 2

t	$w_t(\bar{x}_t-25)$	$w_t(\bar{x}_t-25)^2$	$w_t/\sum w_t$	$(1-w_t(\sum w_t)^2$	$1/f_t(1-w_t/\sum w_t)^2$
1	0.932	2.612	0.130	0.757	0.0398
2	- 1.428	1.285	0.618	0.146	0.0162
3	- 1.817	5.088	0.253	0.558	0.0620
	- 2.313	8.985			0.1180

We have

$$\sum_t s_t(\bar{x}_t-\hat{x})^2 = 8.985-(2.313)^2/2.569 = 6.90 \qquad (31)$$

Substitution into (29) then gives

$$v^2 = \frac{1/2(6.90)}{[1+1/4(0.1180)]} = \frac{3.45}{1.029} = 3.35 \qquad (32)$$

This must be referred to the variance-ratio table entered with $\hat{f}_1 = 2$, and

$$\hat{f}_2 = [3/8(0.1180)]^{-1} = 8/(0.354) = 22.6 \qquad (33)$$

The 5 percent point for F, corresponding to these numbers of degrees of freedom is 3.43. If, therefore, this point is taken as critical, the experiment just fails to demonstrate that the true means μ_t differ.